Some Experiences Developing a Generalized Environmental Data Model

Gerald S. Key

Computer Sciences Corporation
4045 Hancock Street
San Diego, CA 92110-5164 USA
Internet mail: key@cscnet.com

INTRODUCTION

The processes governing the dynamics of marine ecosystems frequently have broad spatial and temporal extents. The resources required to measure these processes are beyond the scope of most monitoring programs. To gain suitable perspectives of these processes requires sharing measurements made by different programs and scientific disciplines for different applications. This in turn requires environmental scientists to report **primary measurement data in fully documented digital form**.

A **primary measurement** is a quantitative observation made in the field or laboratory. It includes *what* was measured, the *quantity* of the measured parameter, and the *units* in which the quantity is expressed. Typical examples of primary measurements include:

12.9 mg/L copper
47 *Acanthurus sandvicensis*
3.6 cm/sec water velocity

Means, standard deviations, diversity indices, and other summary statistics are not a substitute for primary measurement data. They explain less of the variance (i.e., loss of information) and their methods of calculation are subject to change.

A **fully documented** primary measurement also includes supporting information and associated measurements. Supporting information places the measurement in context. It records *where* the measurement was made, *when* it was made, *how* it was made, by *whom*, etc.. Supporting information may have its own additional information requirements. For example, the method used to determine the geographic coordinates, the reference datum, and the significant digits in the coordinate values should be stored with latitude and longitude reported for the measurement location. Associated measurements are made to judge the quality of the primary measurements. Associated measurements might include duplicate and replicate analyses, the detection limit(s) of the analytical method, etc.

Environmental surveys often generate large measurement data sets. Fully documenting these measurements can multiply the size of these data sets significantly. Storing, managing, manipulating, and analyzing this much information is only practical if the data are reported on a **digital** medium. Fortunately, most environmental measurements are today either recorded digitally or converted to digital form for processing. Reporting these measurements digitally is frequently easier than converting them to hardcopy. Converting hardcopy records from earlier

studies to a digital medium is, however, too time-consuming, costly, and error-prone for many purposes.

The Naval Command, Control and Ocean Surveillance Center RDT&E Division (NRaD) in San Diego, California, USA is developing a generalized environmental data model to meet these requirements. NRaD is using this model to implement multi-disciplinary environmental databases, organize and enter historical measurement data into these databases, prepare specifications for reporting environmental measurements, and to share data with other Navy and non-Navy projects. NRaD is also actively pursuing the expansion of a generalized environmental data model to a national or international scale.

BACKGROUND

The Environmental Sciences Division at NRaD is conducting several studies that require sharing measurement data among diverse groups of scientists and resources managers. These projects include an investigation of the sediment quality near the San Diego Naval Station (NavSta) and an integrated environmental compliance program for the Naval Shipyards. Some of the measurement data are being extracted from prior studies and combined with data from on-going studies (including NRaD's) in a common database. The project staffs require broad perspectives of the spatial and temporal distribution of pollutants and other parameters near these facilities. The databases must also support *ad hoc* query and reporting, direct interfaces to statistical, graphical, and other applications, and the extraction of data for use in simulation modeling and other external applications.

While the immediate application of the databases is to support the requirements of the project staffs, NRaD has viewed this effort more broadly. It plans to use the databases as repositories for a variety of other environmental measurement data from prior studies in San Diego Bay and elsewhere. Once organized in this form, NRaD will use the data as a baseline for analyzing the results of future studies. In addition, NRaD is seeking to make the databases and their design available to other organizations. In sum, NRaD required a generalized environmental database rather than another project-specific application.

The design requirements NRaD established for the database design included:

Primary Measurements. The database had to accommodate primary measurements made by different disciplines (e.g., biology, chemistry), in different media (e.g., sediment, water), using different sampling methods (e.g., discrete, continuous).

Full Documentation. The structure of the database should serve as a template for specifying how to report data from external sources for loading into the database. In order to re-use environmental measurements, whether among contemporaneous projects or accumulated as a time-series for future studies, will inevitably lead to unforeseen applications of the data. The re-use of measurements is limited if they lack the information other users require to judge the quality and applicability of the data.

No *a priori* View. In addition to lacking complete documentation, the reuse of environmental measurements is often limited because the data have been organized with an *a priori* view.

Most measurement data are organized into data files, including those created with spreadsheets and statistical programs, rather than into a true database. When a formal database is used, the data are frequently organized to reflect the views of the person who made the measurments: the location where the measurements were made, the date when the measurements were made, the group that sponsored the study, etc. While these are important attributes of environmental measurements, other perspectives are equally valid. The database should not impose a particular perspective on the data. It should instead permit users to reconstruct the perspective of the original investigation from the relationships to other data represented by the model.

Data Quality. The database must prevent unauthorized access to information, whether to view or to change the data. It must ensure the integrity of the database, both the security of the data and configuration control of the database design.

Distributed Data. The days of large, centralized databases have past. For reasons of cost, administration, and use, most environmental data will be managed locally or regionally. Nonetheless, many questions users may want to ask about the data will encompass broader scales. Thus, sharing measurement data will, in many instances, involve linking databases at different locations via a network. The database design must accommodate both central and distributed management of data.

Growth. The database must accommodate change, both in terms of the types of data it stores and the applications that will use those data. The database manager should be able to change the structure of the database without requiring the redesign of the query, reporting, and analytical programs. Similarly, users should be able to upgrade their application software without having to change the structure of the database.

As reported by Michener, *et al*. (1994) and elsewhere (http://www.sdsc.edu/Events/-compeco_workshop/master.html[1]), the environmental sciences are beginning to address the issues of managing measurement data for use beyond the project or application level. However, most of the organized environmental measurement data are still stored in "flat files" rather than databases. Efforts to date have therefore focused primarily on documenting the contents of the data files and transferring data from one file format to another.

Metadata are "data about data." They provide the contextual information about where, when, how, and why the measurements in the accompanying file were made. Metadata files may also record information about how the measurement data were or should be processed. This might include such as algorithms used to convert instrument readouts to standard units or to convert from measurements in one units base to another. In the United States, a standard for geospatial metadata has been developed by the Federal Geographic Data Committee (FGDC, 1994; see also http://corps_geo1.usace.army.mil:80/geo/metadata/mm.03.standard.html).

Geospatial data identify the geographic location and characteristics of natural or constructed features and boundaries on the earth. The growing use of geographic information systems (GISs) for storing and representing spatially distributed data has led to proposed standards in the

[1] This and similar references are Uniform Resource Locators (URLs) to World-Wide Web (WWW) sites on the Internet. Any string of characters in this paper which begins "http://" is a URL.

United States for geospatial data. Two of the major efforts in this regard are the Defense Department's Tri-Service Spatial Data Standards (http://mr2.wes.army.mil/docs/sds.htm) and the National Spatial Data Infrastructure (http://fgdc.er.usgs.gov/nsdi2.html).. Finally, the Ecological Society of America has undertaken an effort to document and record long-term ecological datasets (http://www.sdsc.edu/1/SDSC/Research/Comp_Bio/ESA/FLED/FLED.html).

NRaD has taken a different but compatible approach. Its goal is to eliminate from the storage of information the distinctions between data and metadata, geospatial versus non-geospatial, and other discipline- or application-specific perspectives This approach is driven by the desire to stored fully documented primary measurement data in a database, rather than in data and metadata files. It views metadata simply as other entities about which information needs to be recorded in the same logical structure as the primary measurements they document. Likewise, in this approach temporal, methodological, spatial and other perspectives of these data are equally valid. In short, NRaD is endeavoring to develop a model for organizing environmental measurement data that is generalized in both content and structure. This process of translating the logical representation of an organization's data into a formal structure is termed **data modeling** (Date, 1995). It is a process that is, or should be, independent of the medium used to store the data

Figure 1 shows a portion of the entity-relationship (E-R) diagram NRaD has developed as a generalized data model. An E-R diagram is a logical representation of the *entities* (represented as boxes) about which data are to be stored, the *relationships* (the lines connecting the boxes) between those entities, and the *attributes* (the terms in the boxes) of the entities and their relationships.

Conceptually, an entity (or *relation*) is a two-dimensional table. The columns (*attributes*) of a table define the information recorded about that entity. The relation **tbl_Measurement** in Figure 1 is designed to record information about measurements. It therefore includes attributes that identify what was measured (**Measurement_Parameter_ID**), the quantity of the measured parameter (**Measurement_Value**), the source of the measurement data (**Measurement_Citation_ID**), etc. The attributes above the horizontal line in each entity box in Figure 1 are *primary key* (e.g., **Measurement_ID**). The values of the primary key uniquely identify each row, or *record,* in the table. Attributes below the line are non-keys in that entity. Those designated with "(FK)" are *foreign keys* -- attributes that are non-key in that relation but are part of the primary key in another relation.[2]

Relationships link the primary key in one relation to an appropriate foreign key in another relation. Typically, relationships link the primary key value of a record in the "parent" relation to 0, 1 or many records in the "child" relation. The statement "a sample may include one or more measurements" is equivalent to saying that for each value of the primary key (**Event_ID**) in the **tbl_Sample** relation there may be none, one, or many records of the **tbl_Measurement** relation with the same value in the **Sample_ID(FK)** attribute. This relationship is named **May Include.** It forms the basis for matching information about a sample with information about the measurements performed on that sample. Note that reading the **May Include** relationship in the other

[2] Attributes with the suffix (AK) are *alternate keys.* These fields could be used as the primary key for that entity.

(Many:1) direction is also a true statement: "A measurement can belong to one and only one sample." Relationships may also be 1:1 and (rarely) Many:Many.

RESULTS

The data model depicted in Figure 1 is still under development. It currently includes 60 relationships among 48 entities. Only some of the "core" entities and relationships, those directly related to representing measurements, are shown. Both Figure 1 and the complete data model were generated using ERWin® by LogicWorks, Inc.

An earlier version of the generalized data model was used to implement a database for the NavSta project. This version of the data model, and NavSta database implemented from it, includes 28 relationships among 21 entities. This NavSta database has been implemented from this data model using the Microsoft Access® for Windows relational database management system (RDBMS; see McFadden & Hoffer, 1993). Table 1 summarizes the size and composition of the NavSta database.

Most of the data entered into the NavSta database to date have been extracted from historical sources. Data from measurements made by the project staff, as well as those from other concurrent studies in the Bay, are now being added to the database.

The problems we encountered developing a generalized environmental data model fall into two categories; those related to developing the data model itself, and those related to implementing the data model as an operational database.

Data Model. By far the most difficult problems encountered developing a generalized data model are those related to decomposing natural hierarchies into the flat structure of an entity-relationship model. This problem is illustrated by the natural hierarchy that exists between samples and subsamples (see Figure 2).

One objective of a generalized data model is to make all measurements equally accessible to the user. For instance, the user should be able to search for Parameter = "Copper", without knowing in advance whether the measurement was made on a sediment, water, or tissue sample. Having found a set of measurements, however, the user should be able to retrieve information about the sample medium and any other measurements made on those samples: the size and weight of the fish from which a liver was obtained, a list of other species that were caught in the same trawl, and the measurements that were made on those other organisms.

This recursive association, where one sample belongs to another that belongs to yet another, is represented in the NRaD model as an *unary* ("self-join") relationship (**Is Reallocated Into**) of the **tbl_Sample** entity. In a unary relationship, the primary key (e.g., **Entity_ID**) is also a foreign key (e.g., **Sample_Parent_ID(FK)**)[3] in the same entity. Applying the example data in Figure 2 to this model would result in records something like those in Table 2.

[3] **Sample_Parent_ID(FK)** is an alias, or *rolename*, for **Entity_ID**)

Although this type of representation appears to require complicated record-keeping, in practice it's quite simple. The investigators can use any scheme they choose to assign IDs to samples, as long as they also record the ID of the "parent" sample, if any, from which a sample was derived. This process is recursive to any depth, but the person assigning the IDs only has to be concerned with two levels at a time, the "parent" and the "child."

In situ and *ex situ* measurements are another example of sample hierarchies that present special problems for generalized data models. With an *in situ* measurement, the location and time of the measurement and the site to which the measurement applies are the same. In common parlance, these are "field" measurements. An *ex situ* measurement is made at a place and time that may be far removed from the site where the sample was collected and to which the measurements are meant to apply. The **tbl_Sample_History** entity (not shown in Figure 1) in the NRaD data model accommodates both types of measurements. A record is added to the **tbl_Sample_History** table for every action taken on a sample: "Collected", "Prepared", "Analyzed", "Discarded", etc. The records also include the place and time of the action. Thus, associating measurement data with the action "Collected" will apply them to the field location where the sample was collected. Associating the same measurement data with the action "Analyzed" will apply them to the site where the measurements were made. If the place and time of the "Collected" and "Analyzed" actions are the same for the same sample, the measurements from that sample are by definition *in situ* measurements.

Finally, a generalized data model must be able to represent sample (or other event) geometries that range from points (grab sample), to lines (transects), to areas (quadrats), to volumes (an otter trawl). Simple geometries (e.g., a square quadrat) are relatively easy to represented with two points, such as the longitude and latitude of the upper-left and lower-right corners. Representing irregular geometries requires a more complex data structure. The NRaD data model uses the same upper-left, lower-right algorithm, but permits more than one record for each site. Thus, the coordinates of the largest rectangle that will fit in the sample space are recorded in one row of the **tbl_Location** entity (not shown in Figure 1). If the sample site has a simple geometry, this single record may suffice. If not, the coordinates of the next largest rectangle that will fit in the remaining sample space are added to the **tbl_Location** entity. This process continues, creating records for ever-smaller rectangles, until the entire space has been filled to whatever the desired resolution.

It should be noted that the representation of spatial extent is one of the areas of differentiation between GISs and DBMSs. Most commercial GISs store the geometry of an object (e.g., sample site, road, building, etc.) in a special file format, and store (in the case of the modern client/server GISs) the attribute data associated with that object in a DBMS. This dichotomy between storing graphic objects and their attribute data separately is in part historical, and in part technical. However, as the price/performance ratio continues to decline for both mass storage and processors, the technical reasons for this separation are disappearing.

Database. Without a doubt, the most vexing problem of translating a generalized data model into a generalized database is the issue of missing data. This problem is most acute with historical data. Even when the primary measurement data were reported from prior studies, often only the

parameter, quantity and units were recorded. Location may have to be deduced from a relatively large point on a not-to-scale map; methods may be incompletely described; and quality control information (e.g., calibration data for instruments; voucher specimens for organisms) is rarely present. For the end-user, incompleteness means the measurements have fewer practical applications. For the database administrator, incomplete data can create serious problems when the missing values belong to a primary key that cannot be blank.

A variation on the incomplete data problem is the practice, common in regulatory compliance studies, of reporting "ND" (Not Detected. ND means the measured quantity fell below the Method Detection Limit. Keith (1991) has discussed the arguments for and against this practice from the standpoint of analytical methods. From the perspective of database management, "ND" is a non-value. It cannot be stored in a numeric field, nor can it be replaced by zero or the detection limit. A related problem occurs when the DBMS cannot represent null numeric value (i.e., "no data") as distinct from a value of "0" (a measured value of zero).

Finally, a generalized database may store measurements of the same parameters made by different methods from different sources. This raises the issue of whether to store measurements in their original units or convert them to a common set of units[4]. Storing measurements in their original units is more scientifically pleasing, but can create problems when querying or displaying the data. This is because the numeric quantity (i.e., **Measurement_Value** in Figure 1) will be different for a measurement recorded in g/m^3 than one recorded in lbs/ft^3 -- even though the concentrations of the material in the two samples might be the same. The query "Measurement_Value > 5" thus has no meaning unless these measurements are in common units. Converting measurements to a common units base "on the fly," as the query is being performed, so their magnitudes will be comparable can seriously impact performance. The alternative of converting measurements to common units before loading them into the database carries a price as well. It requires storing the units in which the measurement was originally reported, the significant digits in the original value, and the algorithm used to perform the conversion.

Obviously, many of these difficulties could be avoided if environmental measurements were reported according to a pre-defined specification of the content and format of the data set. NRaD is preparing such a specification, mindful of Slagel's (1994) discussion of the difficulty of developing environmental data reporting standards.

CONCLUSIONS

The investigations reported here have demonstrated it is possible to develop a generalized environmental data model, one that is independent of the discipline that made the measurement and the application that will use the data. These investigations have also demonstrated it is possible to implement this model as an operational, multi-disciplinary database. Such a database can be an effective tool for sharing data among members of the same project, or between projects.

Having demonstrated the feasibility of a generalized environmental data model is only the beginning of the process. The breadth of the data types we have incorporated into our data model is

[4] These arguments apply not only to measurement quantities, but to other units bases as well - such as converting all geographic coordinates to latitude and longitude, converting all local times to Universal Coordinated Time, etc.,

limited when compared to the universe of information that comes under the head of "environmental data.". This is also true of the range of applications to which the resultant databases have been applied. NRaD is continuing to expand both the data model and its application to operational data management systems, and to seek out and work with other organizations with similar objectives. A number of common trends have emerged from these efforts, including:

Generality. Accommodating new types of environmental data in a generalized forces one to decide when the information represents a new entity or simply a new instance of an existing entity. For example, when one group wanted to record information about where and when San Diego Bay had been dredged, we had to decide if "dredging" was a new entity in our model. We eventually decided that "dredging" and "sampling" were both instances of a more common entity, "event". This in turn required rethinking which attributes are common to all events and which ones are specific to certain types of events. This process of aggregating the common attributes into generalized entities occurs with each new type of data we encounter.

Scaleability. The data model must be scaleable in several dimensions. It should, for example, accommodate the kilobytes-to-megabytes of data recorded by field studies in the same organization used for the terabytes of data recorded using remote sensing technology. In theory a measurement is a measurement, and these differences of scale should be deferred to the implementation of the database. In practice the performance issues can cause implementation considerations to percolate back up to the logical data model. Similarly, not every project will need to record information for every entity in the data model. It is important to design the model to identify both the "core" (mandatory) and the optional entities for different classes of applications. An entity such as **tbl_Contract** may be optional for anyone collecting measurement data (because not all measurements are made under a contract), but mandatory for those recording administrative information about environmental data collection. And because the collection and management of environmental data are, and will continue to be, decentralized processes, the data model must be scaleable to support both centralized and distributed database architectures.

Content, Not Format. Another derivative of the requirement for scaleability is the need to implement the data management process on projects with vastly different resources at their disposal. What is possible on a large multi-million dollar project with lots of hardware, software, and technical support may be impractical on much smaller scales. The underlying requirement to obtain fully documented measurements in digital form is, however, the same. The focus should therefore be placed on content -- making sure all the data are recorded in some consistent digital form -- rather than the specific format of the data. The decision to load those data into a large distributed database or stored them on a diskette in a desk drawer is a secondary issue. How we find and use information on networks in the future will probably be vastly different from how we do it today (see Negroponte, 1996). These mechanisms will surely be more forgiving of variable formats than missing data.

Data Independence. How data will be used, whether interactively by a human browsing for information or by an application program, is important in defining what types of information should be included in the data model and recorded in the database. Data management and application development should, however, be separate processes. The database administrator must be able to change the database design without causing the applications that use the database to fail. Likewise, users should be able to change their applications without forcing a design of the data-

base. The concept of independence also carries over into the choice of data management architectures. Data management should be a "server" function, performed in the background and not directly visible to the typical end-user (whether a person or a program). Applications are "client" functions that are used directly the end-user. The client and the server functions may be on the same computer, on different computers at the same site, or on computers that are continents apart. Network technology is making data management distance-independent. The World-Wide Web is a good example of a scaleable, distance-independent client/server technology.

Process Reengineering. There is an old adage in the computer sciences that automating an inefficient process yields an inefficient automated process. Many of the procedures environmental scientists use to make and record their measurements were developed for manual processes. An efficient process with paper and pencil may be inefficient when automated, and vice versa. Data modeling is therefore only part of the picture. Modeling the processes (or functions) that generate or use the data is equally important. Changing, or reengineering, the processes to take advantage of automated information systems is the long-term goal. Hammer & Stanton (1995) discuss process reengineering in the Information Age. How this could be accomplished in something as administratively decentralized as the environmental sciences is an open question. Again, however, the Internet probably provides a good paradigm: set standards as to interoperability, open architectures, and scaleability, then encourage independent development within that framework

REFERENCES

Date, C.J. 1995. *An introduction to database systems*. Addison-Wesley Pub. Co., Reading, Mass.

FGDC. 1994. *Content standards for digital spatial metadata*. Federal Geographic Data Committee. Washington, D.C.

Hammer, M. & S..A. Stanton. 1995. *The reengineering revolution : a handbook*. HarperBusiness, New York.

Keith, L. H. 1991. *Environmental Sampling and Analysis: A Practical Guide*. Lewis Publishers, Chelsea, Michigan.

McFadden, F.R. and J.A. Hoffer. 1993. *Modern Database Management*, 4th edition. The Benjamin/Cummings Publishing., Redwood City, CA.

Michener, W. K., J. W. Brunt and S. G. Stafford (eds). 1994. *Environmental Information Management and Analysis*. Taylor & Francis, London.

Negroponte, N. 1996. *Being Digital.* Vintage Books, New York.

Slagel, R.L. 1994. "Standards for integration of multisource and cross-media environmental data", In: Michener, W. K., J. W. Brunt and S. G. Stafford (eds). 1994. *Environmental Information Management and Analysis*. Taylor & Francis, London.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the assistance Marissa Caballero and Andy Patterson of Computer Sciences Corporation, without whose technical insights and thoughtful suggestions this work would not have been possible. The author also wishes to thank Jeff Grovhoug, Ron Gauthier and Bart Chadwick of the Environmental Sciences Division at NRaD for their encouragement and support. This work was supported by Contract Number N66001-94-D-0090 from the Naval Command, Control and Ocean Surveillance Center.

TABLES & FIGURES

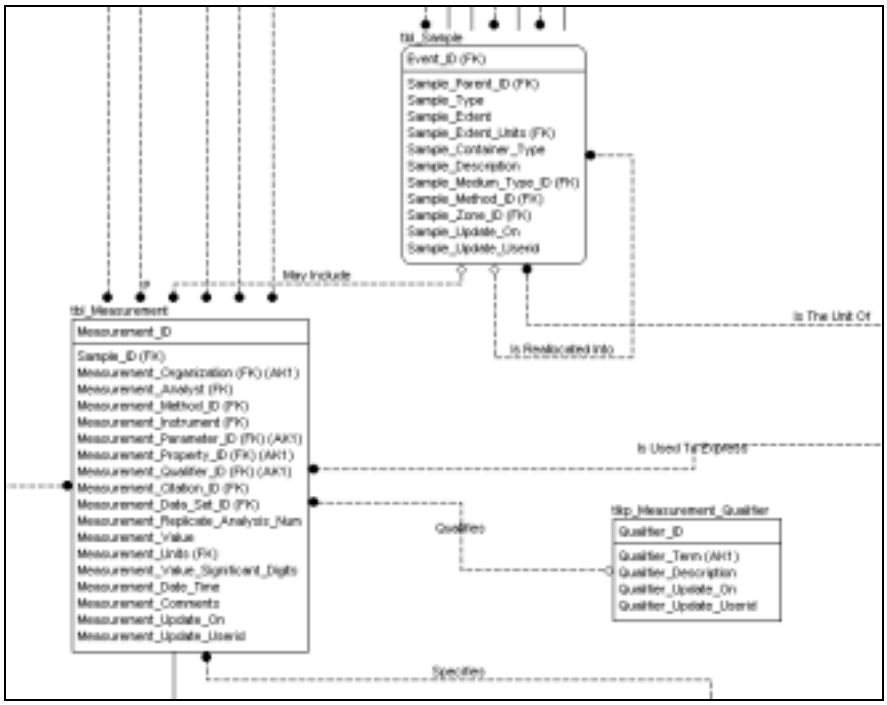

Figure 1. NRaD Environmental Data Model (part)

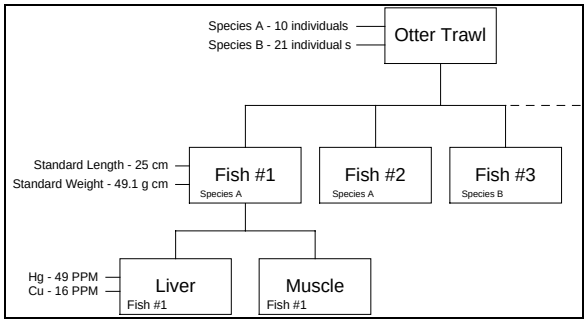

Figure 2. Sample Hierarchy

Number of:	
Measurements	50,456
Samples	1,092
Parameters	256
Data Sources	10
Media	2

Table 1. Summary Statistics for NavSta Database

tbl_Sample

Entity_ID	**Parent_ID**	**…**	**Description**	**…**
12345			Otter Trawl	
67890	12345		Fish #1	
67891	12345		Fish #2	
67892	12345		Fish #3	
78901	67890		Liver	
78902	67890		Muscle	

Table 2. Example of Representing Samples and Subsamples in a Generalized Data Model